Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) DOI: 10.7176/CEIS
Vol.10, No.6, 2019
www.iiste.org

Classification of Under-Resourced Language Documents Using English Ontology

Tsegay Mullu Kassa
Department of Information and Technology, Wachemo University, Hossana, Ethiopia

Yaregal Assabie
Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia

Kidst Ergetie
Department of Information and Technology, Wachemo University, Hossana, Ethiopia

Abstract
Automatic documents classification is an important task due to the rapid growth of the number of electronic documents, which aims automatically assign the document to a predefined category based on its contents. The use of automatic document classification has been plays an important role in information extraction, summarization, text retrieval, question answering, e-mail spam detection, web page content filtering, automatic message routing , etc.Most existing methods and techniques in the field of document classification are keyword based, but due to lack of semantic consideration of this technique, it incurs low performance. In contrast, documents also be classified by taking their semantics using ontology as a knowledge base for classification; however, it is very challenging of building ontology with under-resourced language. Hence, this approach is only limited to resourced language (i.e. English) support. As a result, under-resourced language written documents are not benefited such ontology based classification approach. This paper describes the design of automatic document classification of under-resourced language written documents. In this work, we propose an approach that performs classification of under-resourced language written documents on top of English ontology. We used a bilingual dictionary with Part of Speech feature for word-by-word text translation to enable the classification of document without any language barrier. The design has a concept-mapping component, which uses lexical and semantic features to map the translated sense along the ontology concepts. Beside this, the design also has a categorization component, which determines a category of a given document based on weight of mapped concept. To evaluate the performance of the proposed approach 20-test documents for Amharic and Tigrinya and 15-test document for Afaan Oromo in each news category used. In order to observe the effect of incorporated features (i.e. lemma based index term selection, pre-processing strategies during concept mapping, lexical and semantics based concept mapping) five experimental techniques conducted. The experimental result indicated that the proposed approach with incorporation of all features and components achieved an average F-measure of 92.37%, 86.07% and 88.12% for Amharic, Afaan Oromo and Tigrinya documents respectively.
Keywords: under-resourced language, Multilingual, Documents or text Classification, knowledge base, Ontology based text categorization, multilingual text classification, Ontology.
DOI: 10.7176/CEIS/10-6-02
Publication date:July 31st 2019

1. Introduction

With the rapid growth of the Internet, digital text documents are increasingly replacing the printed ones. With respect to the growth, organizing such “deluge” of documents is a big challenge and organizing them manually is extremely expensive, time consuming, difficult and is often impossible to do. As a result, it originates needs for their automatic classification in order to accelerate and do this hard task.

Automated text categorization, a subfield of NLP, aims at classifying documents to one or more categories [1]. A category represent by a label, and may refer to a class or concept. Recently automatic text classification of digitized documents gained a higher significance due to the rapid growth of digital content. With respect to the growth, organizing them is a big challenge for efficient retrieval of relevant information.

Therefore, finding and improving solutions for text classification has considerable importance. In addition, it extensively used in a wide and diverse range of practical works like spam filtering [2], electronic news classification [3], email classification [4], web page classification [5], and many others.

Many works of automatic text categorization worked based on representative keywords or concepts [6]. In keyword-based approach, a document that categorized should contain a specific keyword that matches the represented document categorized in to the predefined class. This approach spoiled due to some existing practical constraints, mainly existence of vocabulary ambiguities in natural language makes the situation worse and this reduce the accuracy of the classification process. The types of ambiguous terms in natural languages can appear in various forms such as synonyms, homonyms and so on. The incidence of such types of vocabulary ambiguities

DOI: 10.7176/CEIS

poses major challenge in keyword-based classification.

To eliminate such challenge, researchers recommend using an approach that allows classification of documents based meaning rather than keywords and such approach called concept based text categorization. This method extracts concepts from the document and uses those concepts to categorize the document [6]. In order to use concepts to categorize documents, the concepts should represented in the knowledge base, particularly using ontologies. Ontologies recently considered as the de-facto standard for representing semantic information and systematic formalization of concepts, definition, relationships, and rules that captures the semantic content of a domain in a machine-readable format. However, it is difficult to adopt such valuable text categorization approach for under-resource language. Since, language resource is important in order to build ontology, but there is language resource scarceness of such under-resource languages.

In order to attain the ontology based text categorization for under-resource languages without consideration of previously explained difficult, this investigation introduces new techniques that uses ontology of resource rich languages as a knowledge base and perform the classification of documents written in under-resource language.

In general to make the ontology based text classification operable on multilingual environment without language barrier there are two high level approaches in the literature [7]

a) Approach by Translation: - this is an approach, which handles classification of text document with multiple languages over an ontology of a single language using translation. However, the translation of text is a difficult task and is never perfect. To minimize this risk, most approach use translation of term vectors extracted as representative of the given document and not the whole text. This approach provides a modular and flexible framework for addressing such a multilingual text classification problem.

b) Approach by Multilingual Ontology: - this solution is bypasses machine translation in multilingual environments by using a single ontology system to which predetermined manually translated associated concepts in multiple language. To perform mapping of ontology concepts in to different language manually needs human expert and requires more time, more effort and error prone.

In this investigation, to make the proposed approach practicable on a multilingual environment a translation approach adopted. Since, it is suitable for our proposed classification framework because it uses single language ontology as a knowledge base. Beside this, it is modular and flexible approach. For demonstration purpose, the study uses textual document written in under-resource language such as Amharic, Afaan Oromo and Tigrinya.

The remaining part of this paper organized as follows. Section II discusses related works. In Section III, we present the proposed Ethiopian currency recognition system. Experimental results stated in Section IV. Section V presents conclusion and future works.

2. Related Work

The document classification has addressed in server techniques and approaches. In this section, we review existing works in the state-of-the-art on document classification in general, and multilingual text classification in particular.

2.1. Machine Learning Approach to Document Classification

In the past years, few researches had done in text classification for Amharic documents based on Machine-learning approach. However, in this section, we only review the recent work done by [8]. The researcher attempted a hierarchical classification of Amharic news items using support vector machines. The research had conducted with the aim of constructing hierarchical classifier and the experiment had done using a categorical data collected from Ethiopian News Agency (ENA) to evaluate the performance of the hierarchical classifier over the flat classifier. The findings of the experiment show the accuracy of flat classification decreases as the number of classes and documents (features) increase, particularly when the number of top feature set increases. The peak accuracy of the flat classifier was 68.84% when the top 3 features were used. On the other hand, using hierarchical classification show an increasing performance of the classifiers as move down the hierarchy and the maximum accuracy achieved was 90.3% at level -3 (last level) of the category tree, particularly the accuracy increases when the number of top feature set increases as opposed to the flat classifier. The peak accuracy was 89.06% using level three classifier when the top 15 features used. Besides, the performance of flat classifier and hierarchical classifiers compared using the same test data. Thus, it shows that use of the hierarchical structure during classification had resulted in a significant improvement of 29.42 % in exact match precision when compared with a flat classifier.

An automatic text classification for Afaan Oromo documents using machine-learning technique, namely decision tree classifier and support vector machine had been investigated [9]. This study used annotated news texts to train these two classifiers with six news categories i.e. sport, business, politics, health, agriculture and education. In order to preprocess the Afaan Oromo documents, different text preprocessing such as tokenization, stemming, and stop word removal had done. In order to conduct the experiment, 10 fold cross validation technique used. The result of experiment indicated that Decision Tree classifier and Support Vector Machine on six news categories data achieved 96.8 % and 84.93 % respectively. As a result, the researchers concluded that, the machine learning classifiers were applicable for automatic classification of the Afaan Oromo texts.

DOI: 10.7176/CEIS

Beside this, an automatic text classification for Tigrinya text has investigated by only one researcher [10]. This researcher introduced an automatic text classification for Tigrinya text documents with two-step approach. In the first step, clustering used to obtain natural group of the unlabeled data, specifically for this purpose the researcher used direct k-means and repeated bisection clustering algorithm and achieves 0.516 purity, 0.624 entropy and 0.56 purity, 0.611 entropy respectively. Therefore, according to the result the researcher selected the repeated bisection-clustering algorithm to train the text classifier. In the second step, classification performed using Support Vector Machine (SVM) and j48 decision tree classifier. The SVM classifier correctly classified 82.4% with 32.68 seconds whereas the J48 classifier classifies 72% with 34.4 seconds. As a result, the researcher concluded that, the SVM classifier is effective and efficient in classifying the Tigrinya text documents.

2.2. Knowledge Based to Text Classification

A. Kumilachew [8] proposes a framework that automatically categorizes Amharic documents into predefined categories using knowledge represented in the News ontology. The document classification proposed by this paper has three stages. First, all the documents pass through pre-processing stages. Then index terms extracted from a given document mapped onto their corresponding concepts in the ontology. Finally, the selected document classified into a predefined category, based on the weighted concept. The approach was tested and showed that the use of concepts for Amharic document categorizer results in 92.9% accuracy with a promising outcome.

G. Wei et al. [11] propose an ontology based document classifier in order to improve the efficiency and effectiveness of Chinese web document classification and retrieval. The study constructs ontology based Chinese knowledge base named as HowNet and creates ontology for each subclass of the classification system. In this research, RDFS used in order to convert knowledge into ontology and to define the relations among ontology and an ontology relevance-calculating algorithm can classify web documents automatically. The approach tested with SVM, KNN and LSA (TF-IDF) approaches for comparison purpose and according to the experiment result; SVM approach gets average precision rate of 80.1% and the average recall rate of 68.3%. KNN approach gets average precision rate of 82% and the average recall rate of 69.1% and the LSA (TF-IDF) approach gets average precision rate of 82.4% and the average recall rate of 73.8%. The proposed approach achieves experimental result of average

precision rate of 81.9% and average recall rate of 75.8%. The experimental result showed that the proposed ontology based approach achieve highest average precision rate among other three methods (i.e. SVM, KNN, KNN and LSA) and its precision rate most stable.

J. Ma et al. [12] focus on document classification based on the similarities of documents already categorized by ontology using terminology information from the documents. The document classification technique proposed by this paper did not involve any learning processes or experimental data and performed in real time. Their classification results, the precision, recall, and F1 measures are 89.68%, 95.43%, and 92.39% respectively and the F1 measurement compared with TF-IDF and Bayesian method got 79.87% and 82.45%.

2.3. Text Classification and Multilingual

T. Goncalves & P. Quaresma [13] proposed a method to combine different monolingual classifier in order to get a new classifier in which it was suitable for multilingual environment. To build the monolingual classifier, a supervised machine learning approach, particularly Support Vector Machine algorithm used with labeled documents as training data set. The proposed method applied to a corpus of legal documents in four different languages (i.e. English, German, Italian and Portuguese) and evaluated. For testing the proposed method, experiments run over a set of European Union law documents, a set of 2714 full text legal documents. The experiments done using a bag of words representation of documents over SVM algorithm for each language profile and evaluated using a 10 fold stratified cross validation procedure with significance tests done with 90% confidence level. To support the research claim, the experiment conducted for each monolingual classifier and for all possible combiners. According to experimental result, when the Portuguese combiner combines with other language classifier and achieved an average precision of 0.831, when the English and German classifiers are combined resulted in with best average recall of 0.652 and average F1 measure of 0.709. Significance tests show that, for all classes and all performance measures, there was no significance difference between the “best” monolingual classifier and the corresponding combined classifier.

A.Segev. & A. Gal [7] used a lightweight method, which was a design based on multilingual ontology, which means representing the ontological concepts in multiple languages. Therefore, they aim at conveying the local interpretation of ontological concept, thus to overcome the language barrier. In order to analyze the impact of the proposed approach for the support of Multilingual data from news RSS (i.e. which is a format for distributing and gathering content from different sources in different languages across the web) total of 1,778 data items in actual e-Government environment had been used for experiment. In the first experiment, the impact of Multilingual on a single class classification was evaluated in order to analyze the impact of Multilingual on classification recall and on average the use of Multilingual corpus results in a minor reduction of less than 2% in recall (from about 98.3% to 96.58%). On the other hand, in order to analyze the impact of Multilingual on classification precision, it

DOI: 10.7176/CEIS

evaluated on multi class classification. According to experimental result, the precision for all concepts reached 55.17% while the use of multilingual corpora reduced the precision by about 6% to 49.06%. These results indicate that the proposed approach suffers a minor reduction in performance with the introduction of Multilingual.

A.Ferrando et al. [14] used an approach by translation to handle classifying text documents without language barrier over an ontology built on top of a single language. In order to achieve this, the authors used the advantage of BableNet multilingual semantic network, in which words in different languages (specifically all European languages, most Asian languages, and even Latin) grouped into sets of synonyms to cope with multilingual documents. The main goal of the thesis was to provide a modular and flexible framework for addressing such a multilingual text classification problem. In order to demonstrate the potential of the proposed approach the main field of text classification called sentiment analysis or opinion mining implemented.

The experiment for the proposed approach had been conducted over a different threshold value (i.e. Tr = 0 and Tr = 0.2). In the first experiment when the threshold = 0 and in the first variant of SentiModule, smartphones reviews had been used as features ontology to represent the smartphone domain. The experiment results in concerning the reviews with evaluation 5 (i.e. very good) or 4 (good), in case of overall score 5 the approach correctly classifies only the 66.7% of the reviews and similarly only the 60.0% of the reviews with overall score 4. The worst classification results concern the reviews with evaluation 1 (i.e. very bad) and 2 (i.e. bad). Indeed, respectively only in the 43.3% and 38.3% of the cases it produces a correct results. On the other hand, the results of the classification performed using Threshold = 0.2 and concerning the reviews with evaluation 5 (i.e. very good) or 1 (very bad) the approach correctly classifies the 92.6% and 89.1% of the reviews respectively. In case of overall score 4 (i.e. good) or 2 (bad) 88.8% and 82.6% of the reviews was correctly classified respectively.

3. Proposed System

3.1. System Architecture

We propose a new approach solution for classification of under-resourced language written documents with English language ontology without language restriction. As shown in Fig. 1, the proposed approach consists of four major components: pre-processing, translation, concept mapping and classification.

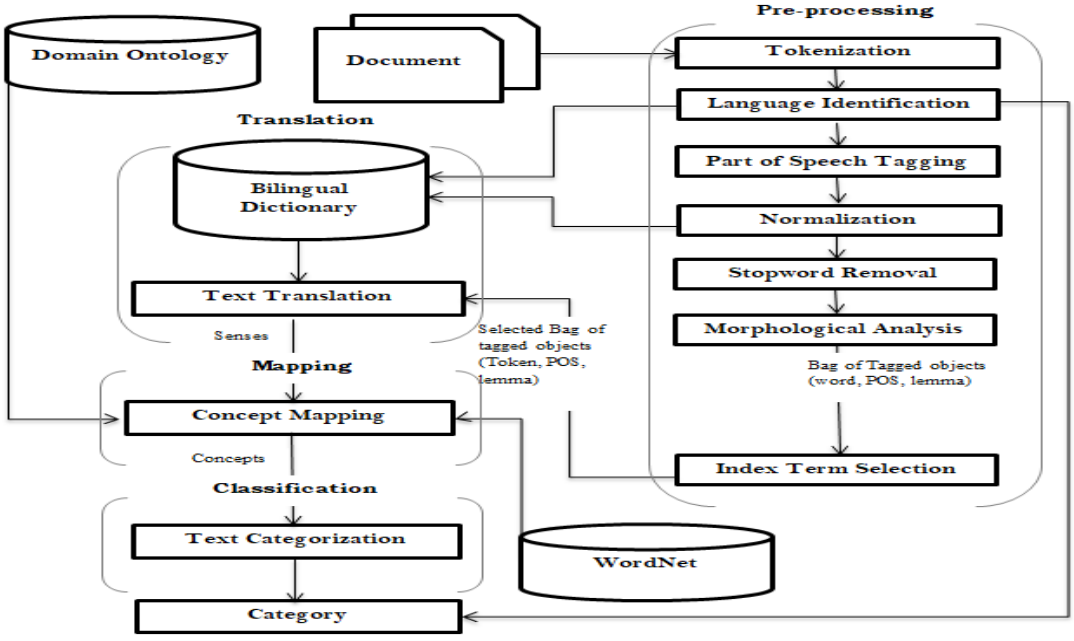

Fig.1. Architecture of the proposed system.

As shown in Fig 1, the input to pre-processing component is a textual document, which classified by the system. The pre-processing component generates a bag of representative tagged objects; each object contains word, Part of Speech and lemma information. The translation component checks the bag of representative tagged objects information along the bilingual dictionary and returns the corresponding bag of sense in a target language. The concept mapping accepts bag of senses and perform mapping of these translated senses along the ontology concepts. During mapping of each translated sense along the ontology concept, the concept weighting computed based on the lemma occurrence in given document. Finally, the classification component accepts bag of weighted concepts and depending on their importance or weight assigned an actual category of a given document.

3.1.1. Pre-processing

The pre-processing component is responsible to accept the input document and produce a set of selected tagged objects after carrying out tokenization, language identification, part of speech tagging, normalization, stop word

removal, morphological analyzer and index term selection. Each selected tagged object contains word, part of speech and lemma information.

a) Tokenization

This sub component is responsible to extract bag of words obtained from the string representation of the input document. In order to extract bag of words from the input document, all occurrences of multiple white spaces replaced by a single white space and the input string split using the word boundary character (i.e. white space character). Finally, after tokenization bag of tokens produced for the next sub-component.

b) Language Identification

This sub component designed in order to recognize the language of the input textual document. This component is backbone of the proposed approach, since the components coming after need to know the language of the input document to perform the classification on multilingual environment. To ensure a fast detection of language, we adopt use of languages stop words as backlist for language identification [15]. Moreover, the languages used for demonstration (i.e. Amharic, Afaan Oromo and Tigrinya) have their own unique stop words and the input for classification is at a document level. As a result, stop word based approach can easily identify the language of input document.

c) Part of Speech Tagging

This sub component involves in determining the POS tag for each words in bag of words provided by tokenization sub component. In this phase due to adoptability and multilingual feature, we adopted TreeTagger [16] and each words are represented by word itself, lemma and part of speech tag (POS). However, due to lack of lemma information, we adopted the hornmorpho [17] in order to generate the lemma of each word. The outcome of this tagger sub component contains each word with POS information. The POS tag of the word used to disambiguate the lemma selection for Morphological Analyzer and to disambiguate the meaning of the word during word-by-word translation. In order to adopt a TreeTagger, we used a training corpus having different POS representation for each language and this would be different POS outcome for each TreeTagger corpora. When a POS outcome is different depending on the text language, it is very complex for our translation task. Hence, in order to handle this complexity we include an operation that converts the original tagged POS labels of each word into the normalized POS labels of the system.

d) Normalization

As shown in Fig 1, this sub component dedicated once the POS tagger was processed. This sub component mainly performs two normalization tasks: The first task is normalization of homophone characters (characters same purpose but different symbol), replacement of such characters with by a common symbol. This reduces the document representation with wrong words. The next task of this sub component is word expansion, which normalize the abbreviated words along with the expanded form of a word.

e) Stop word removal

This sub component dedicated to remove words that deemed irrelevant or non-content bearing words for text classification purpose. In this study, this removal process done automatically by comparing list of words generated for input document along with words in a stop word list. In order to enhance the efficiency of matching words in both collections we used length of words as a feature. From our study, applying stop word removal enhances efficiency of classifier, since it reduces the number of words processed by the upcoming components of the classifier.

f) Morphological Analysis

This sub component is responsible for lemma identification of each words in bag of tagged words in order to each our multilingual document categorization goal. Since, most of the word contents of a bilingual dictionary constructed in lemma form and usually able to translate words correctly in lemma form. In this study, to achieve this we adopt a hornmorpho version 2.5 [17]. To reduce processing time of hornmorpho, we analyze all tagged words in once instead of analyzing a single word at a time. Since, analyzing a single word at a time needs to load model for a target language per each word and this is inefficient.

g) Index Term Selection

In this study, words that have the capability to represent the given document chosen based on their term frequency, occurrence of lemma of a word within actual document. Taking an occurrence of an actual word in order to select an index term is not a good approach, particularly for a language with big morphological complex language like Amharic, Afaan Oromo and Tigrinya. Since, usually equivalent words represented with their inflected word. For example, during term representation of a given Amharic document when "ተማሪ" (student) and "ተማሪዎች" (students) words captured as different words, this degrades performance of index term selection as well the text categorization.

Lemma frequency LF (d, l) is the number of lemma occurrence in the text document and defined as

$$LF\ (di, lk) = \sum_{j=1}^{n} flk \qquad (1)$$

Where d_i is the i^{th} document, l_k is k^{th} lemma of document d_i and $\sum_{j=1}^{n} flk$ is sum of lemma occurrence l_k in a

document di.

3.1.2. Translation

The basic idea behind the proposed document classifier operates on a multilingual environment is due to translation component. This component dedicated to translate representative tagged objects information into English senses. This operation relies on a bilingual dictionary that is in essence mappings between text language tagged objects (i.e. word, POS and lemma) with their corresponding English senses. The bilingual dictionary contains a word in any under-resourced language with their POS and with the equivalent meaning in English language. For efficiency purpose, this dictionary is only loaded for a language of a document after it identified with previous language identification component. In order to enhance the matching probability of terms along the bilingual dictionary during translation, we include normalization like in pre-processing.

To generate the sense from a bilingual dictionary using lemma and POS information is not enough, since bilingual dictionary may not always contain word information in such away. Hence, to enhance the probability of sense generating for all tagged objects, we include four alternative translation functionalities: (i) Find and match a word having lemma and POS information, (ii) Find and match an actual word and POS information, (iii) Find and match a lemma without POS information and (iv) Find and match an actual word without POS information. To get the sense of tagged object all these alternative translation functionalities executed sequentially. Text translation operation is repeated for each word and it is time consuming process, hence to minimize such efficiency issue we employ a technique that perform translation only once for these repeated words in a document.

3.1.3. Concept Mapping

Once the translation generates senses as shown in Fig 1, concept mapping devoted to provide bag of concept for each translated senses. The concept mapping needs ontology as a knowledge base in order to map the translated sense along the ontology concepts. In order to achieve this we use the domain ontology built on top of domain concepts extracted from World News domain ontology (WNO) done by [18]. This component performs lexical and semantic mapping between the ontology concept and the senses provided by the previous translation component.

a) Lexical Mapping

This sub component begins the matching process by calculating string matching between the labels of the class of the ontology with translated senses. The matching process is used along with string pre-processing strategies such as stop word removal, stemming. Such pre-processing tasks used to increase the probability of finding an exact match between the translated senses and the ontology concepts. In this study for stemming tasks, we adopt the Snowball stemmer for English language [19].

b) Semantics based Mapping

There are cases where lexical mapping metrics fail to identify similarity between both parties (i.e. ontology concept and translated sense) that are terminologically different but semantically similar. For example, “student” and “learner” are semantically similar although they are lexically distant from each other. Hence, this problem can be solved by discover semantic relations between them based on various descriptions attached to them. In order to achieve this, we used third part knowledge source called WordNet. As a result, in order to consider semantic relationship between both parties we adopt the combination of three WordNet semantic similarity measures proposed by [14] , which considers WordNet path link (i.e. Shortest path based measure , Wu and Palmer’s Measure) and WordNet information content (i.e. Lin’s measure). Beside this, during concept mapping when compound words found the mapping processed word by word along both parties. On the other hand, like that of translation, mapping is also repeated and time-consuming process. Hence, in this study in order to reduce such inefficiency mapping done only once even if senses appear more than one times.

3.1.4. Classification

It is final component and it is devoted to accept the bag of weighted concepts and assigns one or more concepts based on their importance as a category of input document. When each translated senses mapped along the ontology concepts, a weight assigned to ontology concepts based on mapped lemma occurrence in the input document. In this study, this computed weight feature used to discriminate the importance of mapped concepts return from the ontology. The higher the weight the more important the concept considered to be.

In detail, each concept in ontology O compared with all the translated senses of all the lists S0. Every time when a match found, the label containing the weight of the concept increased by the value associated with the list containing the matching lemma word.

O_w = assignWeights (O) where O_w is the weighted ontology.

To assign weight to an ontology concepts Ci, which mapped for a lemma Lj is the sum of lemma frequency of the jth lemma of a document d.

$$Wci = \sum_{j=1}^{n} f(Lj) \quad (2)$$

Where Wci refers to as weight of concept ci, and f (Lj) is frequency of lemma Lj

Hence, after weighting each ontology concept, our classifier assigns one or more ontology concepts that have a maximum weight as a category of a given text document.

$$Category = Max \begin{cases} i = n \\ Ci \\ i = 1 \end{cases} \quad (3)$$

Where Category refers to assigned document category, Ci is ontology concepts having a weight of greater than zero.

4. Experiment

4.1. Dataset Collection

For evaluating and testing the proposed prototype, different documents collected from each of the news domain specified category used for demonstration of this investigation such as Politics, Business and Economy, Sport, Health, Education and Science and Technology. These testing documents collected from different official sites for all supported under-resource language of the proposed classifier. The Amharic test document for the specified news domain category collected from Fana Broadcasting Amharic program [20] and VOA (Voice of America) [21]. In addition, the Afaan Oromo test document for the specified news domain category collected from Fana Broadcast Afaan Oromo program [20]. On the other hand, for Tigrinya test document for the specified news domain category collected for dimtsi weyane [22] and VOA [23]. On the other hand, in order to evaluate the proposed approach for Amharic and Tigrigna 20 documents used in each news category. As well, for Afaan Oromo 15 documents in each news category are used.

4.2. Implementation

For this work, java used as a development environment. Since, java is pure object oriented programming (OOP) and among benefits of OOP in comparison of other system development that is easy to develop, manipulate, test and understand. OOP clusters things in terms of class and objects so, the testing procedure is undertake by accessing or not accessing different modules according to the given experiment techniques. Toshiba laptop of the following specification used for the research: Windows 10 64-bit operating system. The hardware component comprises of Intel(R) Core(TM) i5 CPU of 2.5 GHz, 4GB memory, and 931GB hard disk.

4.3. Test Results

Evaluation of the proposed classifier done with evaluation parameter that compares the number of documents classified correctly and incorrectly. Typically, the comparison between the documents classified using the proposed classifier and that of the manually classified documents. Beside this, to determine the effectiveness of the classifier we adopt most often-used metrics like precision, recall and F-measure.

In this work, four experimental objectives undertaken to observe the strength of proposed document classifier from different perspectives. Accordingly, we examine the effectiveness of proposed classifier with all features are incorporated and achieved an average of F-measure 92.23%, 89.07% and 88.12% for Amharic, Afaan Oromo and Tigrinya languages respectively. Due to small vocabulary size of compiled bilingual dictionary, the performance decreased for Afaan Oromo and Tigrinya documents. In contrast, we examine effect of morphological analyzer during index term selection and the result of experiment indicates the performance of proposed classifier decrease with 5.38%, 5.19% and 4.8% for Amharic, Afaan Oromo and Tigrinya languages respectively when this sub component is not incorporated. We also examine the proposed classifier without stop word removal and stemming during concept mapping and the experimental result indicates the effectiveness of proposed classifier degrades with an average F-measure of 13.99%, 10.53% and 8.58% for Amharic, Afaan Oromo and Tigrinya respectively. Finally, we examine the proposed approach without semantics based matching during concept mapping and degrade with average F-measure of 13.28%, 16.37% and 10.91% for Amharic, Afaan Oromo and Tigrigna language respectively.

5. Conclusion and Future Works

Build ontology for under-resourced language is very challenging, since it needs domain knowledge and language resource. As a result, ontology based document classification is limited to resourced language (i.e. English) support. We described an approach that automatically classifies under-resourced language written documents with English ontology. Although, to our knowledge, this is the first attempt on under-resourced language written document classification on top of English ontology and we have achieved very promising results that could encourage applying the system in real-world environments. In this work, the number of supported under-resourced language is limited to Amharic, Afaan Oromo and Tigrinya. However, thanks to the system flexibility and modularity it is possible to extend for other languages if bilingual dictionary, POS training corpus and other language specific lexicons are available. Accordingly, future work bound towards improving the performance of the system and set up it in real environment.

REFERENCES

[1] S. International, "Tigrinya," in Tigrinya at Ethnologue, Ethnologue, 2015. [Online]. Available: http://www.ethnologue.com/18/language/tir/. Accessed: Nov. 20, 2016.
[2] A. SinghRathore and D. Roy, "Ontology based web Page Topic identification," International Journal of Computer Applications, vol. 85, no. 6, pp. 35–40, Jan. 2014.
[3] L. Tenenboim, B. Shapira, and P. Shoval, "Ontology based classification of News in an Electronic Newspaper," Intelligent Information and Engineering Systems, Jun. 2008.
[4] K. Taghva, J. Borsack, J. Coombs, A. Condit, S. Lumos and T. Nartker, "Ontology based classification of Email", in International Conference on Information Technology, 2003.
[5] H. Dong, E. Chang and F. Hussain, "An Ontology based Webpage Classification Approach for the Knowledge Grid", in Fifth International Conference ON Semantics, Knowledge, 2009.
[6] M. Sahlemariam, M. Libsie and D. Yacob, "Concept-Based Automatic Amharic Document Categorization", in Americas Conference on Information Systems (AMCIS), 2009.
[7] A. Segev. & A. Gal, Enhance Portability with Multilingual Ontology Based Knowledge Management. Technion: Israel Institute of Technology. Isreal, 2008.
[8] A. Kumilachew, "Hierarchical Amharic News Text Classification", MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia., 2010.
[9] K.Mohammed, R.Babu and Y.Assabie , " Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach", International Journal of Computer Trends and Technology (IJCTT),2017.
[10] G. Assefa, A two-step approach for Tigrinya text categorization. MSC Thesis. Addis Ababa University, Addis Ababa, Ethiopia, 2011.
[11] G. Wei, G. Wu, Y. Gu and Y. Ling, "An Ontology Based Approach for Chinese Web Texts Classification", Information Technology Journal, vol. 7, no. 5, pp. 796-801, 2008.
[12] J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang and O. Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transactions on Systems,
[13] T. Goncalves & P. Quaresma, Multilingual Text Classification through combination of monolingual classifiers. University: Evora, 2010.
[14] A. Ferrando, S. Beux & V. Mascardi, MOo-TC, a Multilingual Ontology Driven Text Classifier. University of Genova, Italy, 2015.
[15]. Truic˘a, C., Velcin, J. and Boicea, A. (2015). Automatic Language Identification for Romance Languages using Stop Words and Diacritics.
[16]G. Salton and M. McGill, *Introduction to modern information retrieval*, 1st ed. Auckland [u.a.]: McGraw-Hill Intern., 1987.
[17] M. Gasser, "HORNMORPHO", 2017.
[18]L.Kallipolitis, V.Karpis, I.Karali, World News Finder: How we Cope without the Semantic Web. Artificial Intelligence and Applications (AIA), IASTED/ACTA Press, 2007.
[19]"Lucene 3.0.3 API", Lucene.apache.org, 2017. [Online]. Available: http://lucene.apache.org/core/3_0_3/api/contrib-snowball/. [Accessed: 24-jun- 2017].
[20]"FBC - እንኪን ወደ ፋና ብሮድካስቲንግ ኮርፖሬት ድረገፅ በደህና መጡ።",
Fanabc.com, 2017. [Online]. Available: http://www.fanabc.com/. [Accessed: 7- Aug- 2017].
[21]"VOA Amharic", ቪኦኤ, 2017. [Online]. Available: https://amharic.voanews.com/. [Accessed: 14- Aug - 2017].
[22]"ድም ወያኔ ትግራይ",2017. [Online]. Available: http://www.dmtsiweyane.com/. [Accessed: 13- Aug - 2017].
[23]"VOA Tigrigna", ቪኦኤ, 2017. [Online]. Available: https://tigrigna.voanews.com/. [Accessed: 21- Aug- 2017].

Authors' Profiles

Tsegay Mullu Kassa received Bachelor Degree in Computer Science from Dilla University, Ethiopia. He received Master's Degree in Information Technology from Jimma University. His research interest includes natural language processing, and Information Retrieval.

Yaregal Assabie received his PhD in Electrical Engineering from Chalmers University of Technology, Gothenburg, Sweden. He received Master's Degree in Information Science and Bachelor Degree in Computer Science from Addis Ababa University, Ethiopia. He is currently working as an Assistant Professor at the Department of Computer Science, Addis Ababa University. His research interests are natural language processing, pattern recognition and digital image processing.

Kidst Ergetie Andargie received Bachelor Degree in Computer Science from Dilla University, Ethiopia. She received Master's Degree in Information Technology from Jimma University. Her research interest includes natural language processing, and Information Retrieval.